Student Evaluations of Teaching (SETs): Limitations in Assessing Teaching Quality and Performance

What do SETs measure and do not measure:

- Accumulating evidence that (SETs) do not measure the quality of teaching or learning outcomes. Randomized controlled experiments have shown that higher ratings are negatively correlated with measures of teaching effectiveness (Stark and Freishtat 2014; Uttl et al. 2017).

- A common issue with past SETs is that students are asked to rate instructors on areas they are ill-equipped to judge. Studies have shown that SETs are associated with student preferences for lower workload, higher grades, type of course, how they enjoyed the course, and students' beliefs about their performance in the class. They are poor correlates of teaching effectiveness or academic progress.

- STEs can be informative to professors and the feedback from students can help professors improve their course. For example, eliciting feedback during the semester, can help professors incorporate the feedback before the end of the semester and emphasize to students that their involvement in the course is valued.

- As a result, there are increasing efforts in many universities to de-emphasize the use of SETs for tenure and promotion decisions. Students voices are important but should not be the primary or only tool for assessing teaching effectiveness.

Gender Bias in SETs:

- In addition to evidence that SETs constitute poor measures for assessing teaching effectiveness, there is also accumulating evidence that they are riddled with bias against women and other underrepresented minorities.

    - Gender bias is typically defined as existing if men and women receive different evaluations that cannot be explained by objective differences in teaching quality (Mengel et al., 2019). This has been documented in both observational studies and experiments.

    - The source of the bias is not clear. It could be related to discrimination, or to students' differential reaction to teaching styles across gender. Disentangling these two mechanisms is difficult and is not required to address the role of bias in SETs in tenure and promotion decisions.

    - Two particularly compelling studies use experimental designs to document gender bias:

        - Mengel et al. (2019) is the first study to use random assignment of students to instructors (some of whom are female and some of whom are male) in a standardized course with identical learning materials and assignments. The study uses a large sample of about 20,000 students. Despite no differences in student performance on standardized tests, grades in future classes, or study effort across male and female instructors, female instructors receive lower scores on teaching evaluations. The magnitude of the bias in not small. In a class with an equal number of male and female students, female instructors receive 14% of a S.D. lower score than male instructors. On a scale from 0 to 1, this implies that female instructors would be ranked on average 0.37 lower than male instructors.

        - MacNell et al (2015) use the setting of an online course, where all students have the same instructor and course materials. However, students are randomly assigned information about the "instructor's" gender. Students who are told they have a male identity instructor give higher scores on the course relative to students who are told they have a female identify instructor.

        - Students own gender also affects this bias. Again, in both observational studies and randomized experiments, the gap in scores between male and female instructors is larger among male students compared to female students.

    - The evidence suggests bias is not only limited to women but extends to other minorities as well. In observational studies, faculty of color and faculty with accents received lower scores than their white counterparts. Experimental evidence suggests bias may exist against LGBTQIA+ instructors as well. However, the evidence on these other types of bias is much more limited, in part because these populations are so underrepresented in academia.

The Risk of Relying on SETs for Personnel Decisions and Possible Solutions

- Relying on SETs as a primary measure for tenure and promotion decisions is coming under increased scrutiny. In fact, the threat from lawsuits (e.g. a class-action lawsuit against universities) is increasing. For example, an arbitrator ruled that Ryerson University in Canada could no longer use student evaluations to gauge teaching effectiveness in promotion-and-tenure decisions.

- A small but growing number of schools have stopped using SETs for personnel decisions, such as tenure, promotion, and merit evaluations. Leading in this effort are USC and the University of Oregon who have stopped using SETs for tenure and promotion decisions. Both universities are moving to a more wholistic approach for evaluating teaching which includes SETs but also rely more heavily on peer reviews, instructors' reflections on own

teaching, eliminating SET questions that rank professors and including more non-open-ended questions to evaluate teaching effectiveness, among other changes.

- CU Boulder, along with the University of Kansas, the University of Massachusetts at Amherst, and other partners are engaged in the Teaching Quality Framework (TQF) initiative with the goal of providing departments with tools to conduct richer evaluations of teaching. According to their website: "The framework draws from multiple sources of evidence of high-quality teaching, including the instructor's materials, peer feedback, and student voices". Importantly, the framework defines teaching as a scholarly activity like research and relies on evidence grounded in scholarship of higher education.

- TQF partners with departments on a voluntary basis and works with faculty to develop evaluations of teaching that are relevant to the context of the department or discipline. Ultimately, the initiative's goal is to generate an approach to teaching effectiveness that is disciplinary-specific yet common to the campus as a whole and is centrally supported. More information on the TQF initiative at CU Boulder can be found here.

- Next steps: Which committee should be actively engaging in this discussion and make recommendations as to the role of SETs in tenure and promotion decisions? Can we invite people involved in the TQF initiative (e.g. Noah Finkelstein from CU Boulder) to discuss the details of the project with the CU Denver community?

We relied on two main sources for this document. The first is a comprehensive literature review of 41 scholarly articles across many disciplines conducted by Holman, Mirya, Ellen Key and Rebecca Kreitzer available here: http://www.rebeccakreitzer.com/bias/ and summarized here: http://www.rebeccakreitzer.com/wp-content/uploads/2019/10/Bias-in-Teaching-Evaluations-Policy-Brief.pdf. The second is the Statement on Student Evaluations of Teaching from the American Sociological Association available here: https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_sept52019.pdf.

Additional Specific Citations are:

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. "What's in a name: Exposing gender bias in student ratings of teaching." *Innovative Higher Education* 40.4 (2015): 291-30.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. "Gender bias in teaching evaluations." *Journal of the European Economic Association* 17.2 (2019): 535-566.

Stark, P. B., and R. Freishtat. "An evaluation of course evaluations." ScienceOpen. *Center for Teaching and Learning, University of California, Berkley. Retrieved from https://www. scienceopen. com/document* (2014).

Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related." *Studies in Educational Evaluation* 54 (2017): 22-42.